Informational requirements of nudging

Benkert, Jean-Michel ; Netzer, Nick

Abstract: A nudge is a paternalistic government intervention that attempts to improve choices by changing the framing of a decision problem. We propose a welfare-theoretic foundation for nudging similar in spirit to the classical revealed preference approach, by investigating a framework in which preferences and mistakes of an agent can be elicited from her choices under different frames. We provide characterizations of the classes of behavioral models in which the information required for nudging can or cannot be deduced from choice data.

DOI: https://doi.org/10.1086/700072

Supplemental Material for: Jean-Michel Benkert, Nick Netzer. 2018. "Informational Requirements of Nudging." Journal of Political Economy 126(6). DOI: 10.1086/700072.

Informational Requirements of Nudging

Jean-Michel Benkert and Nick Netzer

Online Appendix

Supplemental Material for: Jean-Michel Benkert, Nick Netzer. 2018. "Informational Requirements of Nudging." Journal of Political Economy 126(6). DOI: 10.1086/700072.

B Additional Material

B.1 Model Uncertainty

In the main text, we assumed that there is a unique conjecture about the behavioral model, while it may be more appropriate to assume that the regulator considers a number of different models possible. We can replace the assumption of a unique behavioral model by the assumption that the regulator considers any distortion function $d \in D$ possible, where D is a given set of conjectures. For instance, there could be uncertainty about the aspiration level of a satisficer, or one of the models in D could be the rational agent.[1] As a consequence, we no longer have to learn about the welfare preference only, but about the pair $(d, \succeq) \in D \times P$ of the distortion function and the welfare preference.[2]

Let $\bar{\Lambda}(d, \succeq) = \{(d(\succeq, f), f) \mid f \in F\}$ denote the maximal data set generated by the pair $(d, \succeq)$. Then the set of pairs $(d, \succeq)$ that are consistent with data set Λ is $DP(\Lambda) = \{(d, \succeq) \mid \Lambda \subseteq \bar{\Lambda}(d, \succeq)\}$. We again assume that $DP(\Lambda)$ is non-empty, i.e., at least one conjecture is not falsified by the data. Once we have narrowed down the set of model-preference pairs to $DP(\Lambda)$, we obtain the equivalence class of frame f by $[f]_\Lambda = \{f' \mid d(\succeq, f) = d(\succeq, f'),\ \forall (d, \succeq) \in DP(\Lambda)\}$. We can then modify our definition of the binary nudging relation in a natural way, taking into account that both model and welfare preference are unknown. In particular, for any $[f]_\Lambda \neq [f']_\Lambda$ we define $[f]_\Lambda\ N(\Lambda)\ [f']_\Lambda$ if for each $(d, \succeq) \in DP(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$ for all non-empty $S \subseteq X$, so that for each remaining behavioral model the agent's choices under frame f are better than under f', no matter which of the welfare preferences that are consistent with the behavioral model and the data set is the true one.

We are again interested in the existence of an optimal nudge. By the same reasoning as in Section 3.2 of the main text, we consider maximal data sets only. An immediate extension of Definition 2 could require identifiability of $\succeq$ in d, for a given pair $(d, \succeq)$. This property is in fact necessary but no longer sufficient for the existence of an optimal nudge. It rules out that the maximal data set $\bar{\Lambda}(d, \succeq)$ could have been generated by a different welfare preference $\succeq'$ and the same model d, but it does not rule out that it could have been generated by a different welfare preference $\succeq'$ and a different model d'. Since two behaviorally equivalent model-preference pairs $(d, \succeq)$ and $(d', \succeq')$ can have different normative implications (see e.g. Kőszegi and Rabin, 2008; Bernheim, 2009; Masatlioglu et al., 2012), identifiability in the extended setting must aim at all aspects of the pair $(d, \succeq)$ that are normatively relevant.

[1] It is central to the idea of asymmetric paternalism (Camerer et al., 2003) that there are different types of agents, some of which are rational and should not be restricted by regulation.

[2] We continue to assume that there is a non-distorting frame for each pair $(d, \succeq)$, which will typically depend both on the model and on the welfare preference.

Definition 6 *Pair $(d, \succeq)$ is virtually identifiable if for each $(d', \succeq') \in D \times P$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d'(\succeq', f)$.*

Virtual identifiability implies that the welfare preference $\succeq$ is known for sure once the maximal data set has been collected. It still allows for some uncertainty about the behavioral model, but only to the extent that we may not be able to predict the behavior of an agent with a different welfare preference $\succeq' \neq \succeq$.

Proposition 6 *With model uncertainty, $G(\bar{\Lambda}(d, \succeq))$ is non-empty if and only if $(d, \succeq)$ is virtually identifiable.*

Proof. The proof is similar to the proof of Proposition 1 and therefore omitted. ■

We can have multiple models with identifiable preferences each, that, if considered jointly, do not have virtually identifiable model-preference pairs. Model uncertainty of this type poses a fundamental new problem to nudging. On the other hand, adding a rational agent to any given behavioral model with identifiable preferences always preserves the property of virtually identifiable model-preference pairs.

B.2 Imperfectly Observable Frames

In the main text, we assumed that frames are perfectly observable and controllable by the regulator. Since a frame can be very complex, this assumption deserves to be relaxed. The generalization also allows us to model internal states that affect the agent's choices. For instance, consider a satisficing model in which the aspiration level k fluctuates in a non-systematic and unobservable way, as in the original RS model. We can capture this by including the aspiration level into the frame (k affects choice but not welfare), but the extended frame cannot be fully observable and controllable for an outsider.

Imperfect observability can be modelled as a structure $\Phi \subseteq 2^F$ with the property that for each $f \in F$ there exists $\phi \in \Phi$ with $f \in \phi$. The interpretation is that the regulator observes only sets of frames $\phi \in \Phi$ and does not know under which of the frames $f \in \phi$ the agent was acting. The example with a fluctuating aspiration level can be modelled as $F = P \times \{2, \dots, m_X\}$ and $\Phi = \{\phi_p \mid p \in P\}$ for $\phi_p = \{(p, k) \mid k \in \{2, \dots, m_X\}\}$. A behavioral data set is a subset $\Lambda \subseteq P \times \Phi$, where $(\succeq', \phi') \in \Lambda$ means that the agent has been observed behaving according to $\succeq'$ when the frame must have been one of the elements of ϕ'. Thus a welfare preference $\succeq$ is consistent with Λ if for each $(\succeq', \phi') \in \Lambda$ we have $\succeq' = d(\succeq, f')$ for some $f' \in \phi'$, so that $\succeq$ might have generated the data set from the regulator's perspective. The set of welfare preferences that are consistent with Λ is $P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}$, where $\bar{\Lambda}(\succeq) = \{(d(\succeq, f), \phi) \mid f \in \phi \in \Phi\}$ is again the maximal data set for $\succeq$. Note that a non-singleton set of frames ϕ can appear more than once

in a maximal data set, combined with different behavioral preferences. This also implies that the cardinality of $\bar{\Lambda}(\succeq)$ is no longer the same for all $\succeq \in P$, because two different frames $f, f' \in \phi$ might generate two different observations for some preference but only one observation for another preference.

In many applications, such as a satisficing model with fluctuating aspiration level, it is reasonable to assume that the same Φ applies to observing and nudging, i.e., the frame dimensions that the regulator can observe are identical to those that he can control. We allow for the more general case where a set of frames can be chosen as a nudge from a potentially different structure Φ_N.[3] When comparing two elements $\phi, \phi' \in \Phi_N$, we will not necessarily want to compare the agents' choices under each $f \in \phi$ with her choices under each $f' \in \phi'$. For instance, we want to compare orders of presentation for each aspiration level separately, not across aspiration levels. To this end, we introduce a set H of selection functions, which are functions $h : \Phi_N \to F$ with the property that $h(\phi) \in \phi$. The elements of H capture the comparisons that we need to make: when comparing ϕ with ϕ' we compare only the choices under the frames $h(\phi)$ and $h(\phi')$, for each $h \in H$. In the satisficing model we would have one $h_k \in H$ for each aspiration level $k \in \{2, \dots, m_X\}$, defined by $h_k(\phi_p) = (p, k)$. The only assumption that we impose on H is that for each $f \in \phi \in \Phi_N$ there exist $h \in H$ such that $h(\phi) = f$. We can then define the equivalence class $[\phi]_\Lambda = \{\phi' \mid d(\succeq, h(\phi')) = d(\succeq, h(\phi)), \forall (h, \succeq) \in H \times P(\Lambda)\}$ for any Λ and ϕ. As before, for any $[\phi]_\Lambda \neq [\phi']_\Lambda$, let $[\phi]_\Lambda N(\Lambda)[\phi']_\Lambda$ if for each $(h, \succeq) \in H \times P(\Lambda)$ it holds that $c(d(\succeq, h(\phi)), S) \succeq c(d(\succeq, h(\phi')), S)$, for all non-empty $S \subseteq X$.

Let $G(\Lambda) = \{\phi \mid [\phi]_\Lambda N(\Lambda)[\phi']_\Lambda \text{ for all } [\phi']_\Lambda \neq [\phi]_\Lambda\}$ be the set of optimal nudges. We again consider maximal data sets. An immediate extension of Definition 2 could require that for each $\succeq' \neq \succeq$ there exists $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f)$. This property turns out to be necessary but not sufficient for $G(\bar{\Lambda}(\succeq))$ to be non-empty. It implies that the maximal data set for $\succeq$ is different from the maximal data set for every other preference, so that $\succeq$ is identified once $\bar{\Lambda}(\succeq)$ has been collected and once it is known that this set is indeed maximal. Unfortunately, the cardinality of $\bar{\Lambda}(\succeq)$ no longer carries that kind of information, as we could have $\bar{\Lambda}(\succeq) \subset \bar{\Lambda}(\succeq')$ for some $\succeq' \neq \succeq$. Upon observing $\bar{\Lambda}(\succeq)$ we then never know if we have already arrived at the maximal data set for $\succeq$, or if there is an additional observation yet to be made. Our notion of identifiability in the setting with imperfectly observable frames must therefore ensure that the maximal data set reveals itself as maximal.

[3] In continuation of our previous approach, we assume that for each $\succeq \in P$ there exists $\phi \in \Phi_N$ such that $d(\succeq, f) = \succeq$ for all $f \in \phi$. This implies that nudging is not per se impeded by the lack of control over frames. The assumption is clearly much stronger here than before. For instance, it holds in the described satisficing application when there is perfect recall (because the order of presentation that coincides with the welfare preference is non-distorting for all possible aspiration levels) but would not hold with no recall (because the non-distorting order of presentation then depends on the aspiration level).

Definition 7 *Welfare preference $\succeq$ is potentially identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exist $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f')$ for all $f' \in \phi$.*

When frames are not directly observed, identifiability requires more than the existence of a frame $f \in \phi \in \Phi$ that distinguishes between $\succeq$ and $\succeq'$. We can exclude welfare preference $\succeq'$ as a candidate only if the observed distorted preference $d(\succeq, f)$ could not as well have been generated by $\succeq'$ for any other $f' \in \phi$. For instance, no preference is potentially identifiable in the perfect-recall satisficing model with fluctuating aspiration level.[4]

Proposition 7 *With imperfectly observable frames, $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is potentially identifiable.*

Proof. The proof is similar to the proof of Proposition 1 and therefore omitted. ■

We use the term potential identifiability because there is no guarantee that we will ever be able to collect $\bar{\Lambda}(\succeq)$. Even if the agent is exposed repeatedly to a set of frames ϕ, it can still happen that a specific element $f \in \phi$ does not materialize. This is in contrast to the case of observable frames, where a maximal data set can always be collected in exactly m_F steps.

B.3 Complexity

In this appendix, we focus attention on models with identifiable welfare preferences, for which knowledge of an optimal nudge is guaranteed once a maximal data set has been observed. However, collecting a maximal data set requires observing the agent under all m_F frames, which might be beyond the regulator's means. We are thus interested in optimal data gathering procedures and the required quantity of information. The idea is that a regulator, who ultimately seeks to impose the optimal nudge, is also able to impose a specific sequence of frames on the agent, with the goal of eliciting the necessary information efficiently.

For each $t \in \{0, 1, \ldots, m_F\}$, let

$$L_t = \{\Lambda | P(\Lambda) \neq \varnothing \text{ and } |\Lambda| = t\}$$

be the collection of data sets that do not falsify the behavioral model and contain exactly t observations, i.e., observations for t different frames. In particular, $L_0 = \{\varnothing\}$ and L_{m_F}

[4]To see why, note that two preferences which coincide except for the ranking of the two top alternatives are behaviorally equivalent for every order of presentation and every aspiration level $k \geq 2$. This was different if we allowed the agent to be sometimes rational ($k = 1$) as in the original RS model, in which case all preferences are potentially identifiable.

consists of all maximal data sets, i.e., those data sets that the regulator may end up with after observing the agent under all possible frames. Then $L = L_0 \cup L_1 \cup \ldots \cup L_{m_F-1}$ is the collection of all possible data sets except the maximal ones. An elicitation procedure dictates for each of these data sets a yet unobserved frame, under which the agent is to be observed next.

Definition 8 *An elicitation procedure is a mapping $e : L \to F$ with the property that, for each $\Lambda \in L$, there does not exist $(\succeq, f) \in \Lambda$ such that $e(\Lambda) = f$.*

A procedure e starts with the frame $e(\varnothing)$ and, if the welfare preference is $\succeq$, generates the first data set $\Lambda_1(e, \succeq) = \{(d(\succeq, e(\varnothing)), e(\varnothing))\}$. It then dictates the different frame $e(\Lambda_1(e, \succeq))$ and generates a larger data set $\Lambda_2(e, \succeq)$ by adding the resulting observation. This yields a sequence of expanding data sets described recursively by $\Lambda_0(e, \succeq) = \varnothing$ and

$$\Lambda_{t+1}(e, \succeq) = \Lambda_t(e, \succeq) \cup \{(d(\succeq, e(\Lambda_t(e, \succeq))), e(\Lambda_t(e, \succeq)))\},$$

until the maximal data set $\Lambda_{m_F}(e, \succeq) = \bar{\Lambda}(\succeq)$ is reached. Hence all elicitation procedures deliver the same outcome after m_F steps, but typically differ at earlier stages. A procedure does not use any exogenous information about the welfare preference, but the frame to be dictated next can depend on the information generated endogenously by the growing data set.[5]

We now define the complexity n of the nudging problem as the number of steps that the quickest elicitation procedure requires until it identifies an optimal nudge for sure. Formally, let

$$n(e, \succeq) = \min\{t \mid G(\Lambda_t(e, \succeq)) \neq \varnothing\}$$

denote the first step at which e identifies an optimal nudge if the welfare preference is $\succeq$. Since this preference is unknown, e guarantees a result only after $\max_{\succeq \in P} n(e, \succeq)$ steps. With Π denoting the set of all elicitation procedures, we have to be prepared to gather

$$n = \min_{e \in \Pi} \max_{\succeq \in P} n(e, \succeq)$$

data points before we can nudge successfully.

To illustrate the concepts, we first consider the limited search model (assuming $m_X \geq 3$ to make all preferences identifiable). The following result shows that learning and nudging are relatively simple in this model.

[5]Notice that an elicitation procedure dictates frames also for pre-collected data sets that itself never generates. We tolerate this redundancy because otherwise definitions and proofs would become substantially more complicated, at no gain.

Proposition 8 *For any $m_X \geq 3$, the limited search model satisfies*

$$n = \begin{cases} 3 & \text{if } k = m_X/2 \text{ and } k \text{ is odd,} \\ 2 & \text{otherwise.} \end{cases}$$

Proof. We assume $k \leq m_X/2$ throughout the proof, as cases where $k > m_X/2$ can be dealt with equivalently by reversing the role of the first page f and the second page $X \backslash f$ of the search engine.

Case 1: k even. We first construct an elicitation procedure e and then show that it is optimal. Let $e(\varnothing) = f_1$ be an arbitrary subset $f_1 \subseteq X$ with $|f_1| = k$. Now fix any welfare preference $\succeq$. The procedure then generates a data set $\Lambda_1 = \{(\succeq_1, f_1)\} \in L_1$, where $\succeq_1$ agrees with $\succeq$ within the sets f_1 and $X \backslash f_1$. Let a_i denote the alternative ranked at position i within the set f_1 by $\succeq_1$, for each $i = 1, \ldots, k$. Let b_i denote the alternative ranked at position i within the set $X \backslash f_1$ by $\succeq_1$, for each $i = 1, \ldots, k, \ldots, m_X - k$. Then construct the frame $e(\Lambda_1) = f_2$ as $f_2 = \{a_1, \ldots, a_{k/2}, b_{k/2+1}, \ldots, b_k\}$. The procedure then generates a data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\} \in L_2$, where $\succeq_2$ agrees with $\succeq$ within the sets f_2 and $X \backslash f_2$. This construction is applied to all the data sets Λ_1 that are generated by the elicitation procedure for some welfare preference. The elicitation procedure can be continued arbitrarily for all other data sets.

Let $\succeq$ be an arbitrary true welfare preference. We claim that the set $T_k(\succeq)$ of top k alternatives according to $\succeq$ can be deduced from the generated Λ_2, so that the optimal nudge is identified and $n(e, \succeq) \leq 2$ follows. Observe first that none of the alternatives $b_{k+1}, \ldots, b_{m_X - k}$ (if they exist) can belong to $T_k(\succeq)$, because Λ_1 has already revealed that each $b_1, \ldots, b_k$ is preferred by $\succeq$. Now suppose that $b_k \succeq_2 a_1$ holds. We then know that $b_k \succeq a_1$ and thus $T_k(\succeq) = \{b_1, \ldots, b_k\}$. Otherwise, if $a_1 \succeq_2 b_k$ holds, we know that $a_1 \succeq b_k$ and thus $b_k \notin T_k(\succeq)$ but $a_1 \in T_k(\succeq)$. In this case we can repeat the argument for a_2 and b_{k-1}: if $b_{k-1} \succeq_2 a_2$ we know that $b_{k-1} \succeq a_2$ and thus $T_k(\succeq) = \{b_1, \ldots, b_{k-1}, a_1\}$; otherwise, if $a_2 \succeq_2 b_{k-1}$ holds, we know that $a_2 \succeq b_{k-1}$ and thus $b_{k-1} \notin T_k(\succeq)$ but $a_2 \in T_k(\succeq)$. Iteration either reveals $T_k(\succeq)$ or arrives at $a_{k/2} \succeq_2 b_{k/2+1}$, which implies $a_{k/2} \succeq b_{k/2+1}$. In this case, we know that $T_k(\succeq)$ consists of $a_1, \ldots, a_{k/2}$ and those $k/2$ alternatives that $\succeq_2$ and hence $\succeq$ ranks top within $X \backslash f_2$.

Since $\succeq$ was arbitrary, we know that $\max_{\succeq \in P} n(e, \succeq) \leq 2$. Obviously, no single observation ever suffices to deduce $T_k(\succeq)$, neither in the constructed procedure nor in any other one, hence we can conclude that $n = 2$.

Case 2: k odd and $k < m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1, \ldots, a_{(k-1)/2}, b_{(k+1)/2+1}, \ldots, b_k, b_{k+1}\}$, where b_{k+1} exists because $k < m_X/2$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$, in which case $T_k(\succeq)$ consists of $a_1, \ldots, a_{(k-1)/2}$ and those $(k+1)/2$ alternatives that $\succeq_2$ ranks top within $X \backslash f_2$.

Case 3: k odd and $k = m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1, \dots, a_{(k+1)/2}, b_{(k+1)/2+1}, \dots, b_k\}$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$. In this case, we can conclude that $T_k(\succeq)$ consists of $a_1, \dots, a_{(k-1)/2}$, plus either $a_{(k+1)/2}$ or $b_{(k+1)/2}$ but never both, and those $(k-1)/2$ alternatives that $\succeq_2$ ranks top among the remaining ones in $X \backslash f_2$. Hence there exist welfare preferences $\succeq$ for which e does not identify $T_k(\succeq)$ after two steps. Since the missing preference between $a_{(k+1)/2}$ and $b_{(k+1)/2}$ can be learned by having $e(\Lambda_2) = f_3$ satisfy $\{a_{(k+1)/2}, b_{(k+1)/2}\} \subseteq f_3$, we know that $n \leq 3$.

It remains to be shown that $n > 2$. Fix an arbitrary elicitation procedure e and denote $e(\varnothing) = f_1 = \{a_1, \dots, a_k\}$ and $X \backslash f_1 = \{b_1, \dots, b_k\}$, where the numbering of the alternatives is arbitrary but fixed (remember that $k = m_X/2$). Let $\succeq_1$ be the preference given (in ranking notation) by $a_1 \dots a_k\, b_1 \dots b_k$, and consider the data set $\Lambda_1 = \{(\succeq_1, f_1)\}$ and the subsequent frame $e(\Lambda_1) = f_2$. Since k is odd, it follows that at least one of the pairs $\{a_1, b_k\}, \{a_2, b_{k-1}\}, \dots, \{a_k, b_1\}$ must be separated on different pages by f_2, i.e., there exists $l = 1, \dots, k$ such that $a_l \in f_2$ and $b_{k-l+1} \in X \backslash f_2$ or vice versa. Depending on the value of l, we now construct two welfare preferences $\succeq'$ and $\succeq''$. If $l = 1$, let

$$\begin{aligned} \succeq'\colon&\ b_1 \dots b_{k-1}\, b_k\, a_1\, a_2 \dots a_k, \\ \succeq''\colon&\ b_1 \dots b_{k-1}\, a_1\, b_k\, a_2 \dots a_k. \end{aligned}$$

If $l = 2, \dots, k-1$, let

$$\begin{aligned} \succeq'\colon&\ a_1 \dots a_{l-1}\, b_1 \dots b_{k-l}\, b_{k-l+1}\, a_l\, a_{l+1} \dots a_k\, b_{k-l+2} \dots b_k, \\ \succeq''\colon&\ a_1 \dots a_{l-1}\, b_1 \dots b_{k-l}\, a_l\, b_{k-l+1}\, a_{l+1} \dots a_k\, b_{k-l+2} \dots b_k. \end{aligned}$$

If $l = k$, let

$$\begin{aligned} \succeq'\colon&\ a_1 \dots a_{k-1}\, b_1\, a_k\, b_2 \dots b_k, \\ \succeq''\colon&\ a_1 \dots a_{k-1}\, a_k\, b_1\, b_2 \dots b_k. \end{aligned}$$

For the two constructed welfare preferences $\succeq'$ and $\succeq''$, the elicitation procedure first generates the above described data set Λ_1. Subsequently, it generates the same data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\}$, because $\succeq'$ and $\succeq''$ differ only with respect to a_l and b_{k-l+1}, which is not revealed by frame f_2. Since $T_k(\succeq') \neq T_k(\succeq'')$, it follows that $n(e, \succeq') > 2$, which implies $\max_{\succeq \in P} n(e, \succeq) > 2$. Since e was arbitrary, it follows that $n > 2$. ■

The nudging complexity is surprisingly small for the limited search model. In particular, the complexity is not growing in the number of alternatives. This begs the question to what extent the limited search model is representative for more general models. It

obviously always holds that $n \leq m_F$ if all welfare preferences are identifiable, but the number of frames m_F can be extremely large (see footnote 13 in the main text). We therefore derive a tighter bound on n next. The result will rest on the insight that there is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each step. Since there are $m_X!$ different welfare preferences that the agent might have ex ante, an elicitation procedure that reduces the set of possible preferences at each step guarantees identification of the preference and the optimal nudge after at most $m_X! - 1$ steps. It turns out that this bound is tight, because there are models for which it is reached.

Proposition 9 *Any behavioral model with identifiable preferences satisfies $n \leq m_X! - 1$, and there exist models with $n = m_X! - 1$.*

Proof. The result follows immediately if $m_X = 2$. Hence we fix a set X with $m_X \geq 3$ throughout the proof. We denote $m = m_X!$ for convenience.

Step 1. We first derive the upper bound $m_X! - 1$. Consider an arbitrary behavioral model, given by F and d, with $m_F \geq m$ and identifiable preferences. Define

$$\hat{n}(e, \succeq) = \min\{t \mid P(\Lambda_t(e, \succeq)) = \{\succeq\}\}$$

as the first step at which procedure e identifies $\succeq$, and let

$$\hat{n} = \min_{e \in \Pi} \max_{\succeq \in P} \hat{n}(e, \succeq).$$

It follows immediately that $n \leq \hat{n}$, because $P(\Lambda_t(e, \succeq)) = \{\succeq\}$ implies $G(\Lambda_t(e, \succeq)) \neq \varnothing$. We will establish the inequality $\hat{n} < m$.

Consider any e and suppose $\hat{n}(e, \succeq) \geq m$ for some $\succeq \in P$. Since $|P| = m$, there must exist $k \in \{0, 1, \ldots, m-2\}$ such that

$$P(\Lambda_k(e, \succeq)) = P(\Lambda_{k+1}(e, \succeq)).$$

Denoting $e(\Lambda_k(e, \succeq)) = \tilde{f}$ and $d(\succeq, \tilde{f}) = \tilde{\succeq}$, we thus have $\Lambda_{k+1}(e, \succeq) = \Lambda_k(e, \succeq) \cup \{(\tilde{\succeq}, \tilde{f})\}$ and $d(\succeq', \tilde{f}) = \tilde{\succeq}$ for all $\succeq' \in P(\Lambda_k(e, \succeq))$. We now define elicitation procedure e' by letting $e'(\Lambda) = e(\Lambda)$, except for data sets $\Lambda \in L$ that satisfy both $\Lambda_k(e, \succeq) \subseteq \Lambda$ and $f \neq \tilde{f}$ for all $(\succeq, f) \in \Lambda$, which includes $\Lambda = \Lambda_k(e, \succeq)$. For those data sets, we define

$$e'(\Lambda) = \begin{cases} e(\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\}) & \text{if} \quad |\Lambda| \leq m_F - 2, \\ \tilde{f} & \text{if} \quad |\Lambda| = m_F - 1. \end{cases}$$

Note that e' is a well-defined elicitation procedure. First, $\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\} \in L$ holds whenever the first case applies, because $\varnothing \neq P(\Lambda) \subseteq P(\Lambda_k(e, \succeq))$ and Λ does not yet contain an

observation of $\tilde{f}$. Second, the first case then applies repeatedly because $e(\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\}) \neq \tilde{f}$, so that e' only dictates yet unobserved frames.

Consider any $\succeq' \notin P(\Lambda_k(e, \succeq))$, so that $(\succeq_1, f) \in \Lambda_k(e, \succeq')$ and $(\succeq_2, f) \in \Lambda_k(e, \succeq)$ with $\succeq_1 \neq \succeq_2$ for some f. From $\Lambda_k(e, \succeq') \subseteq \Lambda_t(e, \succeq')$ and thus $\Lambda_k(e, \succeq) \nsubseteq \Lambda_t(e, \succeq')$ for all $t \geq k$, it follows that preference $\succeq'$ is unaffected by the modification of the procedure, i.e., $\Lambda_t(e', \succeq') = \Lambda_t(e, \succeq')$ for all $t \in \{0, 1, \ldots, m_F\}$, so that $\hat{n}(e', \succeq') = \hat{n}(e, \succeq')$. Now consider any $\succeq' \in P(\Lambda_k(e, \succeq))$, including $\succeq' = \succeq$. Then $\Lambda_t(e, \succeq) = \Lambda_t(e, \succeq') = \Lambda_t(e', \succeq')$ holds for all $t \leq k$. For $k < t \leq m_F - 1$, the definition of e' implies that $\Lambda_t(e', \succeq')$ does not contain an observation of $\tilde{f}$, and that

$$\Lambda_t(e', \succeq') \cup \{(\tilde{\succeq}, \tilde{f})\} = \Lambda_{t+1}(e, \succeq').$$

Thus

$$P(\Lambda_t(e', \succeq')) = P(\Lambda_t(e', \succeq') \cup \{(\tilde{\succeq}, \tilde{f})\}) = P(\Lambda_{t+1}(e, \succeq')),$$

so that $\hat{n}(e', \succeq') = \hat{n}(e, \succeq') - 1$. Repeated application of this construction allows us to arrive at an elicitation procedure e^* for which $\hat{n}(e^*, \succeq) < m$ for all $\succeq \in P$, which implies that $\hat{n} < m$.

Step 2. We now show that there exist behavioral models with $n = m_X! - 1$, by giving an example of such a model. For any $\succeq \in P$, let $o(\succeq)$ denote the opposite order of $\succeq$. We write $P = \{\succeq_1, \succeq_2, \ldots, \succeq_m\}$, where the numbering of the preferences is arbitrary but fixed. We let $F = P$ and number the frames such that $f_i = o(\succeq_i)$. Finally, let $b : P \to P$ be a bijective mapping such that $b(\succeq) \notin \{\succeq, o(\succeq)\}$, for all $\succeq \in P$. Then the distortion function is given by

$$d(\succeq, f) = \begin{cases} f & \text{if } f \neq o(\succeq), \\ b(\succeq) & \text{if } f = o(\succeq). \end{cases}$$

Note that each frame f_i is non-distorting for a single preference only, the one with which it coincides. This implies $n(e, \succeq) = \hat{n}(e, \succeq)$ for all $e \in \Pi$ and $\succeq \in P$, and thus $n = \hat{n}$. We will establish the equality $\hat{n} = m - 1$.

Consider an arbitrary e. Define i_1 such that $e(\varnothing) = f_{i_1}$, and i_k for $k = 2, 3, \ldots, m$ recursively such that $e(\Lambda_{k-1}) = f_{i_k}$ for the data set

$$\Lambda_{k-1} = \bigcup_{j=1}^{k-1} \{(f_{i_j}, f_{i_j})\}.$$

If $\succeq_{i_m}$ is the welfare preference, then the procedure e will generate the sequence of data sets $\Lambda_t(e, \succeq_{i_m}) = \Lambda_t$ for all $t \in \{0, 1, \ldots, m-1\}$, with $\Lambda_0 = \varnothing$. It follows from the

definition of d that $P(\Lambda_t) = \{\succeq_{i_{t+1}}, \succeq_{i_{t+2}}, \ldots, \succeq_{i_m}\}$ holds for each $t \in \{0, 1, \ldots, m-1\}$. This implies $\hat{n}(e, \succeq_{i_m}) = m-1$, and hence $\max_{\succeq \in P} \hat{n}(e, \succeq) \geq m-1$. Since e was arbitrary, it follows that $\hat{n} \geq m-1$. Together with the result $\hat{n} < m$ established in step 1 of the proof, this implies $\hat{n} = m-1$. ■

The tight bound on n established in Proposition 9 grows more than exponentially in the number of alternatives. This shows that nudging may quickly become infeasible despite the general identifiability of preferences.

References

Bernheim, B. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, 7:267–319.

Camerer, C. F., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *University of Pennsylvania Law Review*, 151:1211–1254.

Kőszegi, B. and Rabin, M. (2008). Revealed mistakes and revealed preferences. In Caplin, A. and Schotter, A., editors, *The Foundations of Positive and Normative Economics*, pages 193–209. Oxford University Press, New York.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. (2012). Revealed attention. *American Economic Review*, 102:2183–2205.